3К

УДК 004.81

Ю.Б. Крапивин
Брестский государственный технический университет, г. Брест, Беларусь
Беларусь, 224017, г. Брест, ул. Московская, 267

Функциональность cross-language в задаче автоматического распознавания семантически эквивалентных фрагментов текстовых документов

Y.B. Krapivin
Brest State Technical University, c. Brest, Belarus
Belarus, 224017, c. Brest, Moskovskaja st., 267

Cross-language Functionality in the Problem of the Automatic Identification of the Semantically Equivalent Fragments of the Text Documents

Ю.Б. Крапивін
Брестський державний технічний університет, м. Брест, Білорусь
Білорусь, 224017, м. Брест, вул. Московська, 267

Функціональність cross-language в задачі автоматичного розпізнавання семантично еквівалентних фрагментів текстових документів

В статье приведен метод решения cross-language-функциональности в задаче автоматического распознавания семантически эквивалентных фрагментов текстовых документов. Данный метод основывается на использовании знаний о естественном языке, затрагивая все его уровни анализа: от лексического до семантического включительно.
Ключевые слова: естественный язык, автоматическая обработка текстов, заимствованный фрагмент.

The article presents the solution of the cross-language functionality in the problem of the automatic identification of the semantically equivalent fragments of the text documents. The solution refers to the using of knowledge of the natural language at all levels of its analysis: from the lexical level up to the semantic one inclusive.
Key words: natural language, automatic text processing, adopted fragment.

У статті наведено метод вирішення cross-language-функціональності в задачі автоматичного розпізнавання семантично еквівалентних фрагментів текстових документів. Даний метод ґрунтується на використанні знань о природній мові, зачіпаючи всі її рівні аналізу: від лексичного до семантичного включно.
Ключові слова: природна мова, автоматична обробка текстів, запозичений фрагмент.

Введение

Информационные системы, оперирующие большими объемами текстовых документов произвольной предметной области и успешно решающие различные прикладные задачи, становятся все более востребованными как предприятиями и организациями, так и отдельными пользователями. При этом обработка информации, представленной в документах на различных языках, в том числе с целью обнаружения семантически эквивалентных фрагментов, не является тривиальной и достаточно актуальна.

Постановка задачи

Обозначенная в названии статьи задача рассматривается здесь в контексте одного важного приложения – автоматического распознавания плагиата, под которым обычно понимают умышленное присвоение авторства на чужое произведение литературы, науки, искусства, изобретение или рационализаторское предложение (полностью или частично). Случаи плагиата могут быть и непреднамеренными, например, вследствие сильного внешнего информационного влияния, которое может проявляться в использовании идей или характерного способа их выражения, а также несоблюдения общепринятых правил цитирования, если речь идет об информации, представленной в текстовой форме [1]. Таким образом, реализацию указанного приложения целесообразно рассматривать в виде последовательности следующих двух этапов:

– распознавание эквивалентных, в определенном смысле, фрагментов у заданного текстового документа и текстовых документов из заданной базы данных и доступных Internet-источников;

– анализ, как правило с привлечением экспертов, эквивалентных фрагментов на предмет их заимствования, т.е. на предмет наличия плагиата.

Если говорить об эквивалентности текстовых фрагментов, то, как показал анализ задачи, речь, в этом смысле, должна идти, конечно же, о полностью совпадающих фрагментах, а также о тех, которые совпадают с точностью до некоторых критериев, определяемых преднамеренными и достаточно нетрудоёмкими действиями (процедурами), предпринимаемыми авторами текстов с целью перевода решения задачи распознавания плагиата из плоскости использования достаточно простых показателей для сравнения текстовых фрагментов в плоскость использования показателей, получаемых на основе серьёзного лингвистического анализа текста, т.е. с целью затруднения решения задачи. К таким процедурам можно отнести следующие:

– перестановка слов, допускаемая с точки зрения грамматики языка;

– (не) использование неинформативных слов, например, вводных конструкций;

– использование синонимов слов для отдельных частей речи (существительных, глаголов, предлогов и т.д.), синонимов залогов и различных синонимических конструкций на уровне именных групп, объектно-параметрических отношений (например, «нагреть А» = «повысить температуру А») и т.п.;

– использование парафраза, т.е. пересказа фрагмента текста, сохраняющего его основной смысл.

Заметим, что последняя из перечисленных процедур основывается, в том числе, и на множестве предшествующих. Что касается основного смысла фрагмента текста, то его можно, например, представить в виде совокупности тех знаний, которые там представлены, рассматриваемых в соответствии с тремя их основными типами [2], [3]: объектами/классами объектов, фактами (семантическими отношениями типа С-А-О, где: С – субъект, А – акция, О – объект) и правилами (причинно-следственными отношениями между самими фактами), отображающими закономерности внешнего мира/предметной области.

Очевидно, что для рассматриваемой задачи, учитывая, что причинно-следственные отношения оперируют фактами, а факты – объектами, можно ограничиться только вторым типом знаний. Таким образом, мы будем говорить, что два текстовых фрагмента являются семантически эквивалентными, если их множества фактов совпадают с точностью до синонимии составляющих их компонентов. И речь, таким образом, идет об автоматическом распознавании в текстовых документах именно таких фрагментов.

Наличие в сети Интернет и полнотекстовых базах данных огромного числа текстовых документов, представленных на различных языках, существенно усложняет качественное решение задачи автоматического обнаружения воспроизведенных фрагментов текстовых документов. Так как требует функциональности cross-language, что в свою черёдь, подразумевает, во-первых, обнаружение в анализируемом, т.е. входном, документе фрагментов из текстовых документов, представленных как на языке этого документа, так и на других языках из рассматриваемого их множества, и, во-вторых, необходимость представления результатов на языке пользователя системы, обеспечивающей решение указанной задачи.

Решение задачи в одноязычной информационной среде

В предыдущем разделе отмечалась актуальность распознавания полностью совпадающих текстовых фрагментов – эта задача была нами решена [1].

Решение же задачи в обобщенной постановке, т.е. распознавание семантически эквивалентных текстовых фрагментов, очевидно, потребует наличия лингвистического процессора (ЛПР), осуществляющего автоматический анализ текста на всех уровнях глубины языка – от лексического до семантического. В качестве такого процессора может быть взят известный многоязычный ЛПР [2].

Текст, практически в любом из используемых ныне форматов (DOC, PDF, RTF, HTML, XML, TXT и др.), поступает на его вход и далее осуществляется преформатирование, лексический (распознавание границ слов и предложений), лексико-грамматический, синтаксический и семантический анализ текста.

На последнем этапе распознаются, в частности, так называемые расширенные факты, т.е. семантические отношения типа SAO (рис. 1, данный процесс иллюстрируется на примере английского языка).

Название компонента	Определение
Subject	*субъект*, концепт выполняющий действие (water is heated by **fire**)
Action	*акция* (действие), выполняемая субъектом над объектом (the workers **build** a house)
Object	*объект*, концепт-получатель действия (**house** is built by the company)
Adjective	атрибут действия – *прилагательное* (the invention is **efficient**; the water becomes **hot**)
Preposition	обстоятельство действия или объекта – *предлог*, обычно в паре с *непрямым объектом* (the lamp is placed **on** the table)
Indirect Object	*непрямой объект* действия, часто в паре с *предлогом* (the lamp is placed on **the table**)
Adverbial	атрибут действия с функцией *наречия* (the object is **slowly** modified; the driver must not turn the steering wheel **in such a manner**)

Рисунок 1 – Структура семантического отношения SAO

Понятно, что при распознавании SAO в конкретных предложениях текстового документа определенные компоненты отношения могут быть пустыми, например, в SAO из предложения «the lamp is placed on the table» компоненты Subject, Adjective и Adverbial не заполняются в силу структуры исходного предложения (рис. 2).

Название компонента	Определение
Subject	-
Action	place
Object	lamp
Adjective	-
Preposition	on
Indirect Object	table
Adverbial	-

Рисунок 2 – Пример неполного семантического отношения SAO из предложения «The lamp is placed on the table»

Очевидно, что знаниям типа «факт» в тексте могут соответствовать разнообразные синтаксические структуры, выражающие, однако, равное либо близкое смысловое содержание. Так, например, факт «fire-heat-water», распознаваемый во фразе «fire heats water», может быть также представлен другими синтаксическими формами:

– water is heated by fire;

– fire is able to heat water;

– using of fire allows to heat water;

– heating of water is accomplished with help of fire.

Дополняя лингвистическую базу знаний указанного ЛПР словарями вводных конструкций и синонимов для отдельных частей речи, определяемых компонентным составом расширенного факта, а его функциональность – соответствующими процедурами поиска по этим словарям, мы тем самым, очевидно, обеспечиваем решение поставленной задачи.

Что касается собственно алгоритма распознавания семантически эквивалентных текстовых фрагментов, то его принципиальная схема аналогична представленному в [1] алгоритму распознавания заимствованных предложений при условии, что текстовый документ рассматривается не как цепочка слов, а как цепочка фактов. При этом могут быть оговорены условия не только полного, но и частичного совпадения таких цепочек как по проценту одинаковых фактов от их общего количества в цепочке, так и по компонентному составу сравниваемых фактов, а также наполнению одинаковых компонентов.

Ниже в качестве примера приводится один из результатов распознавания двух семантически эквивалентных текстовых фрагментов, полученных экспериментальной версией системы.

Фрагмент 1.

…A laser is a device that emits light through a process of optical amplification based on the stimulated emission of photons. A laser consists of a gain medium and optical cavity for providing the optical feedback. The light that is emitted by the laser is notable for its high degree of spatial and temporal coherence…

Фрагмент 2.

…A device that is able to emit light by means of a process of visual amplification that is based on the photons emission is called laser. A gain medium and optical cavity to

provide optical feedback are main parts of laser. The light emitted by the laser is known for high degree of temperature and spatial coherence…

Приведенные текстовые фрагменты состоят из семантически эквивалентных предложений. Так, например, после обработки с помощью ЛПР первых предложений приведенных фрагментов, в них будут выделены соответственно следующие факты:

$F_1^{(1)}$ laser – be – device

$F_2^{(1)}$ laser – emit – light – **through** – process of **optical** amplification

$F_3^{(1)}$ X – base – process of **optical** amplification – on – **stimulated emission of phonons**

$F_1^{(2)}$ laser – be – device

$F_2^{(2)}$ laser – emit – light – **by means of** – process of **visual** amplification

$F_3^{(2)}$ X – base – process of **visual** amplification – on – **photons emission**

Здесь выделены синонимичные компоненты соответствующих фактов: «through» у $F_2^{(1)}$ и «by means of» у $F_2^{(2)}$ и т.д. Непрямые объекты «stimulated emission of photons» ($F_3^{(1)}$) и «photons emission» ($F_3^{(2)}$) признаны синонимичными (условно) в силу принятых в данной версии критериев синонимии именных групп (допускается не учет атрибута). Фиксирование «laser» в качестве субъекта фактов $F_2^{(1)}$, $F_1^{(2)}$ и $F_2^{(2)}$ оказалось возможным благодаря наличию в используемом ЛПР функциональности разрешения анафоры. Знаком «X» в приведенных фактах помечен «пустой» субъект.

Решение задачи в многоязычной информационной среде

Решение задачи автоматического распознавания семантически эквивалентных фрагментов текстовых документов в многоязычной информационной среде требует, очевидно, организации, во-первых, распознавания языка текстового документа и, во-вторых, машинного перевода (МП) текстов во множестве L заданных языков, $L=\{L_i\}$, $i=\overline{1,n}$. Причём, в последнем случае речь может идти о разработке / использовании либо множества систем МП с языка L_i на язык L_j, $i,j=\overline{1,n}$, $i \neq j$ (случай, когда все языки из их множества L являются «функционально равными»), либо множества систем МП с L_i на L_j, $1 \le j \le n$ – фиксированное, $i=\overline{1,n}$, $i \neq j$ (случай, когда один из языков из множества L, а именно L_j, является «функционально базовым»). Такой подход к организации машинного перевода текстов имеет место, если существуют многоязычные системы МП, осуществляющие качественный перевод тестовых документов во множестве заданных языков. Причем, выбор «функционально базового» языка позволяет оптимизировать решение задачи, как по трудоемкости, так и по скорости: в случае, если язык входного документа совпадает с языком, выбранным в качестве базового, то документ подвергается немедленной обработке, иначе – предварительно переводится на базовый язык системой МП. В этой постановке задача была нами решена для текстов на русском и белорусском языках [1], но разработанные при этом алгоритмы пригодны для многих языков.

Как показал проведенный анализ, на данный момент существующие системы МП не обеспечивают приемлемых результатов работы для решения указанной задачи в общем случае, в связи с чем предлагается идея использования языка-посредника – «интерлингвы», который будет являться «функционально базовым языком». Его основу могут составить уникальные семантические понятия – концепты и факты, которые в принципе от языка не зависят. Что касается формы его представления, то наиболее удачной, по нашему мнению, является структура многоязычной лексической БД MModWN, включающей множество двуязычных словарей, достаточных для обеспечения качествен-

ного перевода фактов, полученных с помощью указанного выше ЛПР. MModWN, аналогичная по своей структуре WordNet [4], описывает концепты внешнего мира в форме пронумерованных понятий (синсетов), выраженных набором синонимичных слов и словосочетаний на всех языках из множества L, а также различными семантическими отношениями между концептами («общее-частное», «часть-целое», «группа-элемент» и т.д.) [5].

Таким образом, в силу вышеизложенного имеет место следующая схема системы автоматического распознавания воспроизведенных фрагментов текстовых документов, реализующей cross-language-функциональность, представленная на рис. 3.

В соответствии с представленной структурно-функциональной схемой для каждого документа, будь то документ из Полнотекстовой базы данных, содержащей множество эталонных документов, базы данных релевантных Интернет-доступных документов, полученных в результате Интернет-поиска, или входной документ, заданный пользователем, определяется язык его представления в Подсистеме определения языка текстового документа. Затем документ обрабатывается в Подсистеме автоматического индексирования документов, в которой для каждого документа, строится его поисковый образ (ПОД) – множество фактов, полученных с помощью ЛПР, и, используя возможности элемента лингвистической базы знаний (ЛБЗ) – лексической БД MModWN, в свою очередь позволяющей осуществлять его перевод на ЕЯ из поддерживаемого множества, наряду с оригинальным документом сохраняется в поисковый индекс – Проиндексированные входной и из полнотекстовой БД документы или Проиндексированные Интернет-доступные документы, если документ был получен в результате Интернет-поиска по ключевым словам, выделенным из анализируемых документов. Далее подключается функциональность Подсистемы поиска релевантных документов, которая реализуется путём сравнения их ПОД-ов.

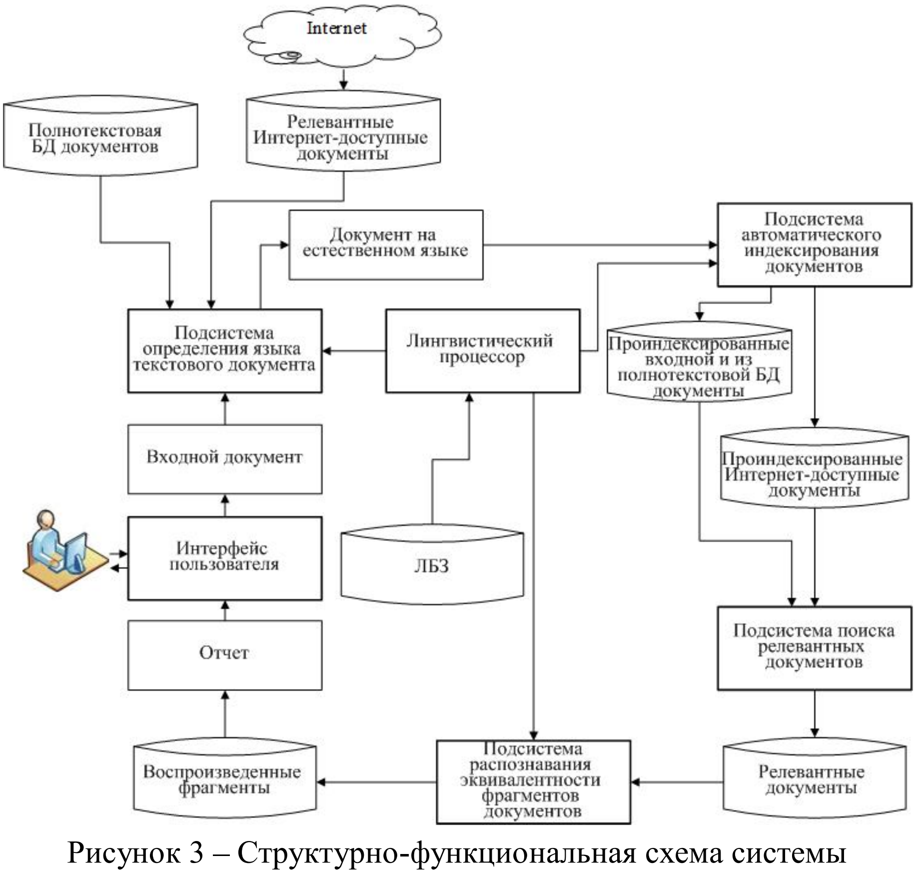


Рисунок 3 – Структурно-функциональная схема системы

На следующем шаге входной документ и все полученные для него релевантные документы поступают в Подсистему распознавания эквивалентности фрагментов документов, которое осуществляется с учётом явного и указанного ранее (например, по компонентному составу сравниваемых фактов) типа неявного заимствований. Эквивалентные с точки зрения критериев системы фрагменты – Воспроизведенные фрагменты – с указанием их источников оформляются в виде Отчёта и поступают пользователю. Его взаимодействие с системой осуществляется посредством интерфейса, который поддерживает ввод документов и просмотр результатов поиска заимствований, приведенных на языке источника.

Функциональность Подсистемы определения языка текстового документа, Подсистемы автоматического индексирования документов, Подсистемы распознавания эквивалентности фрагментов документов обеспечивается ЛПР и его ЛБЗ, причём в той мере, в какой это необходимо для качественного решения задачи, то есть, как отмечалось ранее, с учётом семантического уровня языков.

Выводы

Представленный в работе метод решения cross-language-функциональности в задаче автоматического распознавания семантически эквивалентных фрагментов текстовых документов основан на использовании языка-посредника, обеспечиваемого специальной многоязычной лексической базой данных и существенно расширяющего возможности существующих инструментально-программных средств анализа текстовых документов на предмет выявления в них случаев заимствования без ссылок на авторов.

Литература

1. Крапивин Ю.Б. Автоматический поиск заимствованных из Интернет-источников фрагментов // Искусственный интеллект. – 2012. – № 4. – С. 183-189.
2. Совпель И.В. Система автоматического извлечения знаний из текста и её приложения // Искусственный интеллект. – 2004. – № 3. – С. 668-677.
3. Совпель И.В. Автоматическое распознавание причинно-следственных отношений в текстовых документах // Искусственный интеллект. – 2005. – № 4. – С. 646-650.
4. WordNet [Электронный ресурс]. – 2013. – Режим доступа : http://wordnet.princeton.edu/ – Дата доступа : 17.04.2013.
5. Постаногов Д.Ю. К вопросу многоязычности систем инженерии знаний и их приложений / Д.Ю. Постаногов, И.В. Совпель // Искусственный интеллект. – 2006. – № 3. – С. 474-479.

Literatura

1. Krapivin Y.B. Avtomaticheskij poisk zaimstvovannyh iz Internet-istochnikov fragmentov // Iskusstvennyj intellect. – 2012. – № 4. – S. 183-189.
2. Sovpel I.V. Sistema avtomaticheskogo izvlechenija znanij iz teksta i ejo prilojenija // Iskusstvennyj intellect. – 2004. – № 3. – S. 668-677.
3. Sovpel I.V. Avtomaticheskoje raspoznavanije prichinno-sledstvennyh otnoshenij v tekstovyh dokumentah // Iskusstvennyj intellect. – 2005. – № 4. – S. 646-650.
4. WordNet [Elektronnyj resurs]. – 2013. – Rejim dostupa : http://wordnet.princeton.edu/ – Data dostupa : 17.04.2013.
5. Postanogov D.Y., Sovpel I.V. K voprosu mnogojazychnosti system injenerii znanij i ih prilojenij // Iskusstvennyj intellect. – 2006. – № 3. – S. 474-479.

RESUME

Y.B. Krapivin

Cross-language Functionality in the Problem of the Automatic Identification of the Semantically Equivalent Fragments of the Text Documents

The article presents a method for the cross-language-functionality in the problem of the automatic recognition of the semantically equivalent fragments of the text documents, which is considered in the context of one important application – automatic plagiarism identification. It implies, at first, the detection in the input document the fragments of the text documents presented in both the language of the document and the other languages of their considered set, and secondly, the need to present the results in the language of the user of the system that ensures the solution of the problem mentioned above. The usage of the intermediate language – «interlingua», which contains the unique semantic notions – concepts and facts, which does not depend on the language in principle, and provided with special multilingual lexical database are suggested.

Thus, the method presented in the article is based on the usage of the knowledge of the natural language and refers to the analysis of the text in all depth levels of language: from lexical up to semantic one inclusive. It significantly extends the capabilities of the existing software tools of analysis of the text documents with the purpose of recognition in them of the adoptions without citing the authors.

Статья поступила в редакцию 10.04.2013.